DOI: 10.1093/nar/gkh076

DBTSS, DataBase of Transcriptional Start Sites: progress report 2004

Yutaka Suzuki[1],*, Riu Yamashita[1,2], Sumio Sugano[1] and Kenta Nakai[1,2]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan, and [2]Undergraduate Program for Bioinformatics and Systems Biology, Faculty of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

Received September 15, 2003; Revised and Accepted October 1, 2003

DDBJ/EMBL/GenBank accession nos+

ABSTRACT

DBTSS (http://dbtss.hgc.jp) was originally constructed based on a collection of experimentally determined TSSs of human genes. Since its first release in 2002, it has been updated several times. First, the amount of stored data has increased significantly: e.g. the number of clones that match both the RefSeq mRNA set and the genome sequence has increased from 111 382 to 190 964, now covering 11 234 genes. Second, the positions of SNPs in dbSNP were displayed on the upstream regions of contained human genes. Third, DBTSS now covers other species such as mouse and the human malaria parasite. It will become a central database containing data for many more species with oligo-capping and related methods. Lastly, the database now serves for comparative promoter analyses: in the current version, comparative views of potentially orthologous promoters from human and mouse are presented with an additional function of searching potential transcription-factor binding sites, which are either conserved or diverged between species.

INTRODUCTION

The knowledge of exact transcriptional start sites (TSSs) of genes is valuable in many ways: it makes the prediction of translational start sites more accurate; it can be used for exploring sequence determinants of TSSs; and it makes the analysis of upstream regulatory regions (promoters) more precise. In principle, information of a TSS is obtained by mapping the corresponding transcript onto the genome sequence. Nevertheless, it is widely known that many mRNA sequence data stored in public databases, lack information about their 5′ ends because of the difficulty in obtaining full-length cDNAs. Thus, even after the completion of human genome sequencing, it is not easy to locate TSSs systematically. To overcome this problem, we have developed a method to construct full-length enriched cDNA libraries using a cap selection technique, the oligo-capping method, and have been systematically collecting full-length cDNA data with this method [(1); T.Ota *et al.* submitted]. Initial computational characterization of human TSSs has been carried out (2,3) and a database [DataBase of Transcriptional Start Sites (DBTSS)] containing the TSS information of 7889 human genes has been constructed (4). In this report, we summarize the updates of DBTSS since its first release, including its new departure as a basis of comparative promoter analyses.

NEW FEATURES

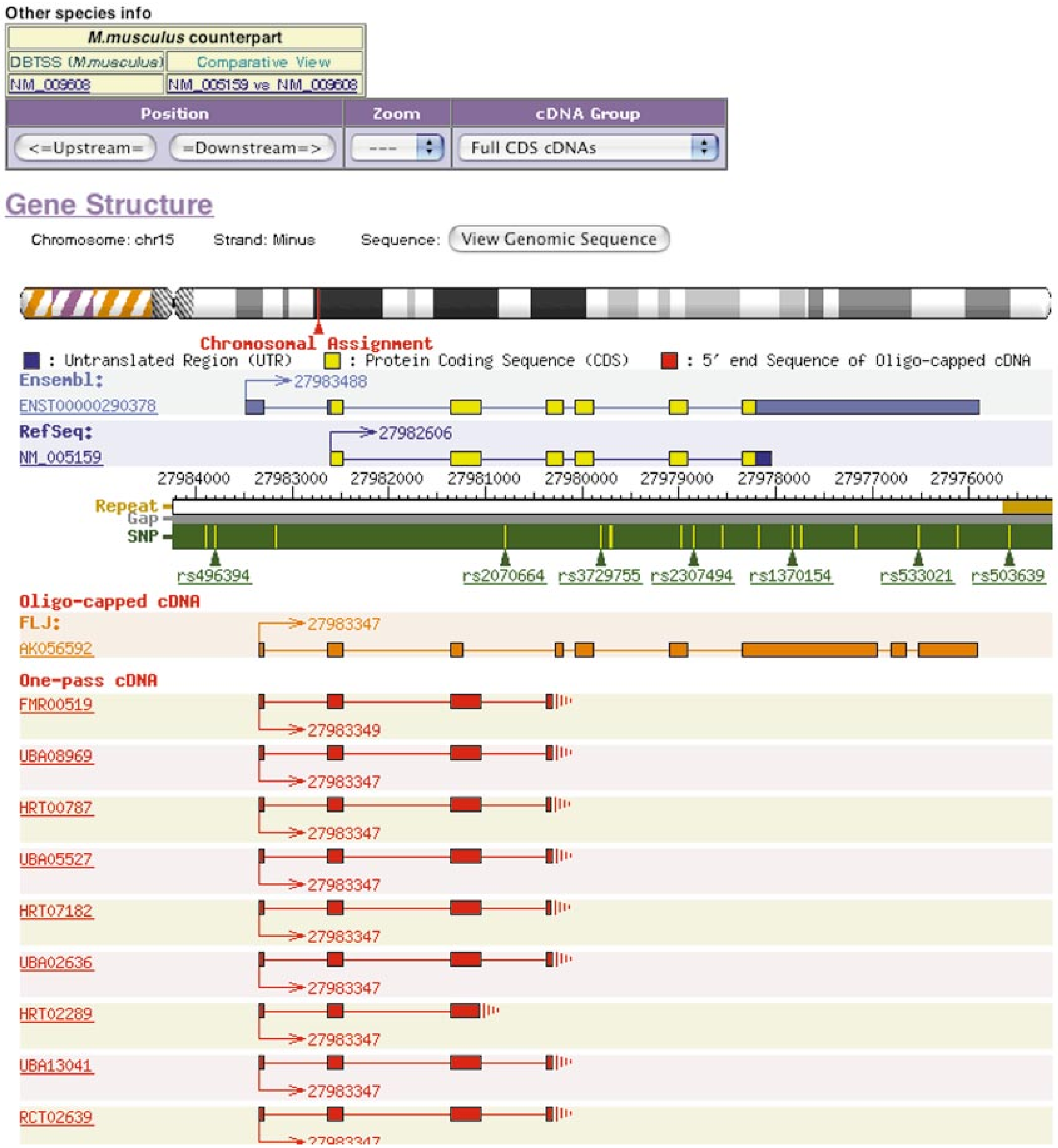
Compared with its initial version, the current DBTSS (version 3) has been upgraded in at least five ways. First, the number of processed one-pass human cDNA clones has increased significantly (from 217 402 to 400 225). Since one of the important findings from our TSS analysis was that the TSS position of a gene is not always fixed but rather often fluctuates for ~50 bp on average (3), the distribution of TSS positions should become clearer as the number of mapped cDNA clones increases. As always, we constructed a so-called RefFull sequence set (11 234 sequences) by extending the 5′-end sequences of RefSeq mRNA sequences (5), if necessary. On average, 6042 sequences were extended by 71.6 bp. At the genomic level, the average difference between 5′-ends of two data sets becomes 4396 bp because of internal introns. Thus, it is clear that our data make promoter analysis of human genes much easier. For more details of the statistics of the DBTSS, see the Statistics section of the DBTSS web page.

Second, to facilitate promoter analysis of human genes, we mapped the positions of single nucleotide polymorphisms (SNPs) stored in a public database, dbSNP (5), on the –1000:+200 region of each representative TSS for each human gene (a sample output is shown in Fig. 1). These SNPs are candidates of functional regulatory SNPs (rSNPs) that affect the promoter activity. We also plan to add SNP data from other sources. In DBTSS, it is also possible to enlist the name of genes located within a specified distance from each SNP.

The third, and probably the most important, upgrade of DBTSS is that it now supports data from multiple species. To date, we have constructed many full-length cDNA libraries of various species upon requests from many researchers. In addition, large-scale collections of cDNAs determined using a related method by Yoshihide Hayashizaki's group are also publicly available (6,7). In the current version, we added the

*To whom correspondence should be addressed. Tel: +81 3 5449 5343; Fax: +81 3 5449 5416; Email: ysuzuki@ims.u-tokyo.ac.jp

+BP192706–BP383670



Other species info

M.musculus counterpart	
DBTSS (*M.musculus*)	Comparative View
NM_009608	NM_005159 vs NM_009608

Position	Zoom	cDNA Group
<=Upstream= =Downstream=>	---	Full CDS cDNAs

Gene Structure

Chromosome: chr15 Strand: Minus Sequence: View Genomic Sequence

Figure 1. Example of the output of a human gene including the correspondence with a mouse gene, gene position in the chromosome, comparison with Ensembl and RefSeq data, SNP positions and graphical representations of one-path cDNA clones.

data of 2490 clones of *Plasmodium falciparum*, the human malaria parasite (8) and 580 209 full-length cDNA sequences of *Mus musculus* (7). The number of Ref-full members for mouse is 6875 (for more details, see Y.Suzuki *et al.*, submitted). We will add data for other species whenever we get the agreement. They include data for *Caenorhabditis elegans*, chimpanzee, macaque, *Cyanidioschyzon melorae* (unicellular red alga), zebrafish and sorghum.

The remaining two novel features will be explained in the next section.

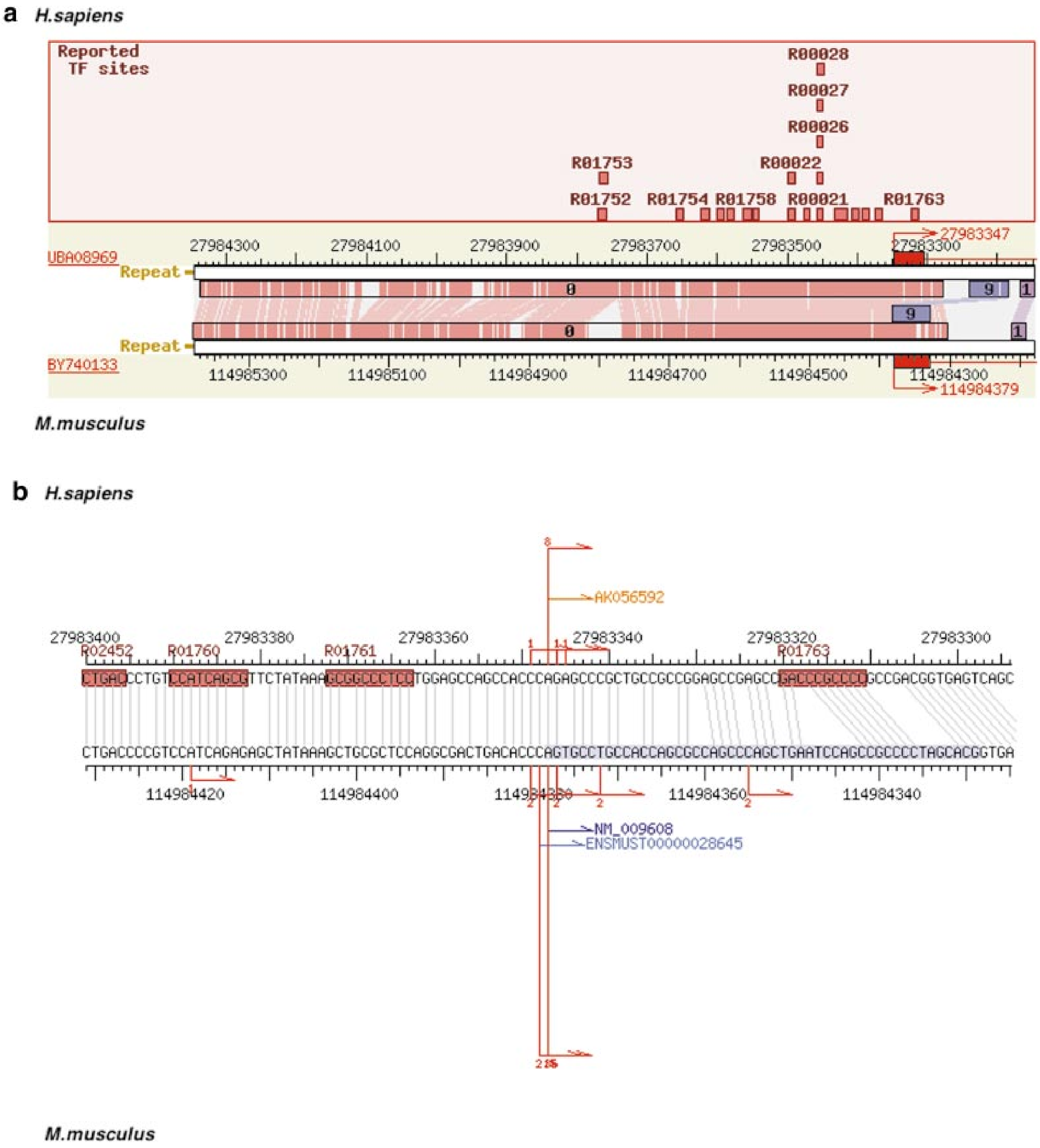


Figure 2. A comparative view of human and mouse promoters. (**a**) Global view with potential transcription factor binding sites. Locally similar sequence segments are shown in boxes and the corresponding boxes are represented by the same number (e.g. '0'). (**b**) More detailed view around the corresponding TSSs.

PROMOTER COMPARISON AND SEARCH OF *CIS*-ELEMENTS

The fourth novel feature of the DBTSS (version 3) is that it provides users with comparative views of human and mouse promoters that are probably orthologous. The potentially orthologous gene set was obtained from the LocusLink database (5) and our own sequence comparison. As a result, promoters of 3324 gene pairs can now be displayed. In each pair, locally similar sequence segments were detected by a local alignment program, LALIGN (9) and their correspondences are shown graphically (Fig. 2).

The fifth novel feature is a function for locating positions similar to known transcription-factor binding sites, which are stored in the TRANSFAC database (10). More specifically, we support TRANSFAC Public-based search (for searches using TRANSFAC Professional, which is a commercial version, users should follow its condition of use, which are shown in our web page). To reduce the number of potentially spurious hits, users can choose various levels of cut-off values and target regions/strands. Moreover, it is also possible to restrict hits within conserved regions between the two species. It is also possible for users to enlist gene names that specify combinations of the above conditions: e.g. genes that harbor both potential binding sites of factors A and B on their upstream regions could be selected with arbitrary cut-off values. With this function, the DBTSS can now be regarded as a platform of systematic promoter analyses.

DBTSS is available at http://dbtss.hgc.jp/ and will continue to expand, incorporating our in-house data and others.

ACKNOWLEDGEMENTS

We thank T. Hasui, K. Abe, M. Morinaga, M. Ishizawa, M. Kawamura, T. Mizuno, A. Kanai and H. Hata for technical support; J. Mizushima-Sugano and E. Nakajima for helpful discussion; Y. Hayashizaki for permission to incorporate their mouse data into DBTSS; and E. Wingender and A. Kel for enabling TRANSFAC-based search. This study was supported by a Grant-in-Aid for Scientific Research on Priority Areas and by special coordination funds for promoting science and technology (SCF), both from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

REFERENCES

1. Suzuki,Y. and Sugano,S. (2003) Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.*, **221**, 73–91.
2. Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Nakamura,Y. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.
3. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Morishita,S. *et al.* (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.*, **2**, 388–393.
4. Suzuki,Y. Yamashita,R., Nakai,K. and Sugano S. (2002) DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
5. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
6. Carninci,P. and Hayashizaki,Y. (1999) High-efficiency full-length cDNA cloning. *Methods Enzymol.*, **303**, 19–44.
7. The FANTOM consortium and the RIKEN Genome Exploration Research Group Phase I & II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
8. Watanabe,J., Sasaki,M., Suzuki,Y. and Sugano,S. (2002) Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. *Gene*, **291**, 105–113.
9. Huang,X.Q., Hardison,R.C. and Miller,W. (1990) A space-efficient algorithm for local similarities. *Comput. Appl. Biosci.*, **16**, 373–381.
10. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A. E,, Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.